DOCUMENT RESUME

ED 139 821 TM 006 310

AUTHOR Moyer, Judith E.; Fishbein, Ronald L.
TITLE A Comparison of Kuder-Richardson Formula 20 and Kappa as Estimates of the Reliability of Criterion-Referenced Tests.
PUB DATE [Apr 77]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Variance; Comparative Analysis; *Criterion Referenced Tests; Measurement Techniques; *Statistical Analysis; *Test Reliability
IDENTIFIERS Kappa Reliability Index; Kuder Richardson Formula 20; Parallel Forms Reliability

ABSTRACT

The problem that this research addressed was one of decision making. Given three sets of criterion-referenced tests which were designed to be parallel in content, would a traditional reliability coefficient produce different decisions about the reliability of those tests than would kappa? The procedure used collected statewide results on 136 test pairs. Each test in a pair consisted of five items; and each test measured a performance objective that was considered to be minimal by curriculum and measurement specialists. Three samples of test pairs were selected: one represented the range of the parallel forms definition of reliability; one the Kuder-Richardson Formula 20 (KR-20) definition; and one sample represented the range of kappa. Parallel forms, KR-20, and kappa were the three levels of the independent variable. The two dependent variables were the mean KR-20 for each item pair and the kappa for each item pair. One analysis of variance was performed to find out if a difference existed among the mean KR-20's for the three groups; a second analysis of variance was performed to find out if a difference existed among the mean kappas for the three groups. No significant difference was found among the KR-20's for the three groups. A significant difference was found among the mean kappas for the three groups. The Tukey Post Hoc method of group mean comparison revealed that both the KR-20 and the kappa group differed from the parallel forms group, but they did not differ significantly from each other. The obtained value of kappa does seem to be related to the homogeneity of the items in a test, as measured by KR-20, but it was not concluded that KR-20 could be in any way predicted from kappa, at least based on this study. (Author/MV)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes eve effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects t quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDR EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made fr

ED139821

A COMPARISON OF KUDER-RICHARDSON FORMULA 20 AND KAPPA AS ESTIMATES OF THE RELIABILITY OF CRITERION-REFERENCED TESTS

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Judith E. Moyer, Michigan Department of Education

Ronald L. Fishbein, United States Office of Education

TM006 310

*This paper was written by Ronald L. Fishbein in his private capacity. No official support or endorsement by the USOE is intended or should be inferred.

The problem that this research addressed was one of decision making. Given three sets of criterion-referenced tests which were designed to be parallel in content, would a traditional reliability coefficient produce different decisions about the reliability of those tests than would kappa.

The procedures used collected statewide results on 136 test pairs. Each test in a pair consisted of five items; and each test measured a performance objective that was considered to be minimal by curriculum and measurement specialists.

Three samples of test pairs were selected: one represented the range of the parallel forms definition of reliability; one the Kuder-Richardson Formula 20 (KR-20) definition; and one sample represented the range of kappa. Parallel forms, KR-20, and kappa were the three levels of the independent variable. The two dependent variables were the mean KR-20 for each item pair and the kappa for each item pair. One ANOVA was performed to find out if a difference existed among the mean KR-20's for the three groups; a second ANOVA was performed to find out if a difference existed among the mean kappas for the three groups.

No significant difference (.05 level) was found among the KR-20's for the three groups. A significant difference (beyond the .05 level) was found among the mean kappas for the three groups. The Tukey Post Hoc method of group mean comparisons revealed that both the KR-20 and the kappa group differed from the parallel forms group, but they did not differ significantly from each other. The obtained value of kappa does seem to be related to the homogeneity of the items in a test, as measured by KR-20, but it was not concluded that KR-20 could be in any way predicted from kappa, at least based on this study.

Reliability of measurement is of primary importance in any scientific endeavor, and educational testing is no exception. Since the early 1900's, researchers have been developing theory and methods for estimating reliability. For any kind of measure, be it a scale, a ruler or a test, "reliability" has been accepted to mean "consistency." Stanley (1971), however, points out that operationalizing the concept of "reliability" or "consistency" is not simply a process of selecting and using an available formula,

> There is no single, universal and absolute reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem. The appropriate allocation of variance from different sources calls for practical judgment of what use is to be made of the resulting statistical value. (p. 363)

This issue has become particularly clear with the wide scale use of criterion referenced tests. Some knowledgeable researchers question the use of traditional methods of estimating reliability with criterion referenced tests. They believe that the conditions which must be logically satisfied in order to use the traditional methods cannot be satisfied. As will be discussed in more detail later, the possible lack of variation among the scores on criterion referenced tests suggests that traditional reliability estimates which largely depend on variation may be inappropriate. However, this position is not accepted by all researchers. Various educational researchers have maintained that traditional methods are appropriate with criterion referenced tests (Ebel, 1975; Livingston, 1972), while others have maintained that the nature and use of criterion referenced tests demand reinterpretation of traditional methods (Hambleton and Novick, 1973), or new methods altogether (Swaminathan, Hambleton and Algina, 1974). Clearly, there is a need for empirical data to clarify the nature of these theoretical positions.

The test-retest method of estimating reliability, wherein examinees are administered a test at one time and are subsequently administered the same test at another time, would seem to be the most logical and definitive means of operationalizing the concept of reliability. If individuals received the same scores on both tests, one could assume the measure to be reliable. This would be analogous to measuring a board with a ruler two times. If the results were the same both times, the carpenter could consider the measure to be a reliable one. However, students of human behavior have a problem in establishing the reliability of their measuring devices by this means that is not shared by all scientific researchers: that is the problem of reactivity of human subjects to measurement. The human, unlike a length of board, for instance, often changes as a result of being measured, so that subsequent measurements may be different, not necessarily because of errors in the measuring instrument, but because of changes that have occurred over time in the individuals being measured. For this reason, and other practical considerations, such as time and money, other methods of estimating test-retest reliability have been developed and are frequently used. Two commonly reported methods are parallel forms and KR-20.

Parallel forms, although they are used by commercial test publishers, suffer from the extreme difficulty of their development. Parallel forms must meet the following criteria: they must have equal mean item difficulties, equal item difficulty variances, equal item variances, equal variances of item variances, equal means of covariances within and between tests as has been well explained by Horst, (1966). The application of the theory, as might be expected, in practice suffers in comparison to the prescription (Horst, 1966; Bohrnstedt, 1970). However, Bohrnstedt maintains that "roughly parallel" forms can be developed and Stanley (p. 405, 1971) says that preparing

"parallel forms should not present undue difficulty." Apparently the question to be answered is "what should these parallel forms uniformly do?" If the purpose of the parallel tests is to rank each individual in the same order relative to others, then this should be possible to accomplish, even if the theoretical criteria are impossible to satisfy in their entirety.

Kuder-Richardson Formula 20 (KR-20), the second estimate of reliability, measures the internal consistency, homogeneity, of the items of a test by treating each item as a parallel test with every other item. Horst (p.262, 1966) points out that homogeneity, although it contributes to the estimation of reliability, is not, in and of itself, a measure of reliability, because it does not "indicate solely the extent to which the measures yielded by the individual items can be relied upon."

Obviously, these traditional methods of estimating the reliability of traditional, norm referenced tests are subject to some operational problems, but the introduction of criterion referenced tests has heightened educational researchers' awareness of the old problems as well as directed attention to possible new problems. Norm referenced tests are expected to produce variability, to spread examinees along a continuum according to each individual's level of the quality being measured, therefore, most reliability measures depend on the variability of the scores. Criterion referenced tests, however, are designed to measure whether or not examinees have attained a specific skill. Usually, they are short tests, administered shortly after a period of instruction to determine which students have attained a skill and which have not. For this reason, variability among students is expected to be low, possibly nonexistent, if all the students have mastered the skill.

Theoretically, then, as Stanley (1971, p. 435) points out, "a criterion referenced test can give reliable information even though its classically defined reliability coefficient equals zero." In other words if one could

use the test-retest method in this real situation i.e., if the problem of reactivity were absent, the measure would be found to be reliable. In this case, the operationalization of the concept of "consistency" with a classical reliability coefficient would be, logically, a very poor choice. Again, it should be remembered that since criterion referenced tests are usually short, operational use of test-retest method of determining reliability would also be a poor choice in that the memory effect, the phenomenon of examinees remembering their previous answers on the retest, could produce spuriously high reliability coefficients.

Consistent with the use of criterion referenced tests for classifying students into mastery and non-mastery categories on a specific skill, Millman (1974), Hambleton and Novick (1973) and Swaminathan, et. al., (1974) have recommended that the consistency of decisions regarding these two categories be the proper operationalization of the concept of reliability. The consistency of decisions across two tests is measured by kappa, and these researchers urge that this consistency of the decision process, as measured by kappa, is a reflection of the quality of the content and the use of the test for making mastery-non-mastery decisions.

In practice, little is known about kappa, but it appears to provide an appropriate operational definition of reliability if one accepts the notion that criterion referenced tests are for classification of students into mastery and non-mastery categories. The coefficient of agreement indicates the proportion of agreement about masters and non-masters beyond what would have been expected by chance across two (or more) versions of a test.

This research attempted to provide some answers to the question: if these methods, traditional and new, were used with data obtained from administering criterion-referenced tests, would different decisions about the reliability of those tests be made on the basis of the two reliability estimates, Kuder-Richardson Formula 20 and Kappa.

Methods

Data Source

The criterion referenced tests which were the basis for this study were the 4th and 7th grade mathematics tests from the 1975 Michigan Educational Assessment Program (MEAP). Each test consisted of five items which measured a minimal performance objective. Attainment of an objective is defined as at least four correct responses; non-attainment is defined as fewer than four correct responses. Both the tests and the objectives were written by Michigan mathematics educators and specialists.

Every September since 1973, all Michigan fourth and seventh graders, except specified special education students, take the MEAP tests. During the 1975 program, experimental tests measuring the same objectives were tried out on a systematic random sample of students for the purpose of developing a bank of new test items.

Each sample of students responded to a few "new' tests as well as the entire regular MEAP test on the same answer sheet. By the sampling method used, a sample of about 4500 students per five item test was obtained.

Analysis

The units of analysis for this study were the test pairs obtained from the sampled students' responses to the regular test of an objective and an experimental test of an objective. In all, there were 136 test pairs.

All of the 136 pairs of tests were ranked from "very high" to "acceptable" on each of the three criteria for selection, parallel forms, mean KR-20's and kappa. They were then stratified into four strata of reliability: very high, high, moderate, and acceptable. From each stratum five pairs of tests were selected to be the sample on which the analysis was performed. The assumption behind this process was that the test develop-

ment procedures were such that all of the tests would have at least "acceptable" reliability. ("Acceptable," as specifically applied to each of the levels of the independent variable will be discussed below, since no general definition was appropriate.) If, in some cases, this assumption of acceptable reliability proved to be unfounded, unacceptable pairs were discarded. In Table 1 are shown the three levels of the independent variable and the strata from which the samples of test pairs were selected

Table 1

Sampling Plan for Stratified Random Selection of Test Pairs That Represent Characteristics of Parallel Forms, Mean KR-20 and Kappa

Reliability Coefficients	Level 1 Parallel Forms Test Forms	Level 2 Mean KR-20 Test Forms	Level 3 Kappa Test Forms
Very High	5	5	5
High	5	5	5
Moderate	5	5	5
Acceptable	5	5	5
TOTAL	20	20	20

Parallel forms sample. The first sample to be drawn was that which represented the range of parallel forms. As stated before, each triplet of test forms that measured an objective was written to be parallel in content. However, the traditional, norm-referenced meaning of "parallel" is much more specific than human judgment about content; therefore, the following requirements for parallel forms were the basis for selection:

1. Each test had the same number of items (five).
2. The mean item difficulties were equal.
3. The dispersions of item difficulties were equal.
4. The variances of the item variances were equal.
5. The means of the inter-test item covariances were the same between tests as within tests.
6. The variances of the distribution of the covariances were the same between tests as within tests.

Since all of the tests consisted of five items, all test pairs were retained using the first criterion. The second step consisted of ranking all parallel forms on the basis of equal mean item difficulties (p-values). If p-values were within .02 points of each other (with rounding from the third decimal place), they were operationally defined to be "equal." These pairs were ranked on their mean p-values and were judged to be in the Very High stratum if the mean p-values differed between the two test forms by only .000 to .004; the High stratum if mean p-values differed from .005-.009; the Moderate stratum if mean p-values differed from .010-.014; and the Acceptable stratum if mean p-values differed from .015-.025.

The rationale for carrying the stringent definition of parallel forms only to the second criterion consisted of two main points. 1) the test pairs had been designed to be parallel in content and 2) an operational definition of "equal" for the remaining criteria would be an arbitrary one with little empirical evidence on which to base it. In this study, therefore, the obtained values for the other criteria were accepted to be "equal."

In all three samples, parallel forms, mean KR-20, and kappa sample, the number of cases which fell into a stratum was determined and each case was assigned an identification number, then a random number table was used to select the cases which would represent the stratum in the sample. For instance, if there were 18 members in the stratum, all numbers between 01

and 18 were eligible for selection. The cards on which the data were recorded were shuffled; then the random number table was used by the researcher so that the first five numbers between 01 and 18 which were encountered indicated the cards which were to be selected for the stratum of a given sample.

KR-20 sample. The second sample to be drawn was the sample representing the range of mean KR-20's. The mean KR-20 was obtained by adding together the KR-20 for the core test and the KR-20 for the paired experimental test and then dividing the sum by two (the number of test forms). The mean KR-20 was used in order to produce a coefficient which would be mutual to the test form pair, the unit of analysis. The Very High mean KR-20 stratum was defined as those pairs having a mean KR-20 of .85 or higher; the High stratum consisted of those pairs having a mean KR-20 of .79-.84; the Moderate stratum consisted of those having a KR-20 of .73-.78; and the Acceptable stratum consisted of those with a mean KR-20 of .60-.72. In addition, if the KR-20's from the two tests differed by more than .10, they were not considered for inclusion in the sample, since they seemed to differ so much from one another that their values, taken separately, would have clearly put each one into a different stratum from the other.

The kappa sample. After the parallel forms sample and the KR-20 sample were selected, the kappa sample was randomly selected from the remaining test pairs, again according to the stratification plan outlined above. The remaining pairs were ranked from high to low according to the obtained values of kappa. Since kappa's obtained value is strongly affected by the marginal proportions, it is difficult to determine the range of specific values of kappa which may be obtained from the data. Kappa will reach 1 only if the decision making is perfect across both forms of the tests. Therefore, it was determined that a kappa of .30 define

the lower limit on the Acceptable stratum. This would indicate, after the effects of chance agreement were removed, the proportion of joint agreements across the test. Intuitively, this seemed to be the lowest "acceptable reliability" that could be tolerated. Further, it was decided that the other three strata be defined as .40-.49, .50-.59, and 60-1.00. Because of a lack of empirical evidence on the behavior of kappa, this plan seemed to be reasonable.

The dependent variables. Using the test pairs, the mean KR-20's and kappas were calculated. These two reliability coefficients were the dependent variables. Kappa, the coefficient which measures the consistency of decision making across two measures, produced one coefficient for each test pair. Obtaining the mean KR-20 for each test pair, however, involved two steps. 1) The KR-20 for each five-item test was calculated. This measured the homogeneity of the five items within each test. 2) The mean KR-20 for each test pair was calculated by adding together the KR-20 computed for each of the tests in the test pair and then dividing by two (the number of tests). This was done in order to have one KR-20 measure for each test pair, the unit of analysis

Since the question to be answered by this study was whether KR-20 and kappa would give different estimates of the reliability of these criterion-referenced tests, two analyses of variance were planned. The independent variable in each analysis was the definition of reliability. One group was composed of test pairs of "acceptable" reliability as defined by Horst's parallel forms criteria, the second group was acceptably reliable as defined by KR-20, and the third, as defined by "acceptable" kappas. Each

group consisted of 20 test pairs. For the first ANOVA, the dependent variable was the mean KR-20 for the three groups; for the second ANOVA, the dependent variable was the mean kappa.

Since the focus of this research was on the possible differences in estimation of reliability that the three methods might produce, an analysis of variance was performed on the KR-20 coefficients and the kappa coefficients, the dependent variables, for each of the three groups. The null and research hypotheses to be tested were:

$$H_o: \quad \mu_{1,KR-20} = \mu_{2,KR-20} = \mu_{3,KR-20} \tag{1}$$

$$H_1: \quad \sum_{j=1}^{3} (\mu_{j,KR-20} - \mu_{\cdot KR-20})^2 \neq 0 \tag{2}$$

where 1 = level of the independent variable representing the range of parallel forms

2 = level of the independent variable representing the range of mean KR-20

3 = level of the independent variable representing the range of kappa

$$H_o: \quad \mu_{1,\kappa} = \mu_{2,\kappa} = \mu_{3,\kappa} \tag{3}$$

$$H_1: \quad \sum_{j=1}^{3} (\mu_{j,\kappa} - \mu_{\cdot\kappa})^2 \neq 0 \tag{4}$$

where 1 = level of the independent variable representing the range of parallel forms

2 = level of the independent variable representing the range of mean KR-20

3 = level of the independent variable representing the range of kappa

Results

In Table 2 are shown the results of the ANOVA which addressed the question of whether or not there was a difference of the mean KR-20's for the three groups. As can be readily observed, the differences between the mean KR-20's for the three groups were not found to be significant at the .05 level. The three group means and standard deviations are shown in Table 3.

In Table 4 are shown the results of the ANOVA which addressed the question of whether or not there was a difference of the mean kappas for the three groups. As shown in the Table, there was a difference which was significant at the .05 level. The three group means and standard deviations are shown in Table 5.

Table 2

Results of the ANOVA to Test the Hypothesis that the Means of the Mean KR-20's for Three Groups Were Unequal

Source	Sum of Squares	df	Mean Square	F	Probability
Between	.02	2	.01037	1.311	.28
Within	.45	57	.0079		
Total	.47	59			

Table 3

The Means and Standard Deviations of the Mean KR-20 for Each Level of the Independent Variable

	Size	Mean	Standard Deviation
1, Parallel Forms	20	.740	.0888
2, KR-20	20	.786	.0780
3, Kappa	20	.761	.0987

Table 4

Results of the ANOVA to Test the Hypothesis that the Mean Kappas for Three Groups Were Unequal

Source	Sum of Squares	df	Mean Square	F	Probability
Between	.158	2	.07899	5.728	.0054
Within	.786	57	.01379		
Total	.944	59			

Table 5

The Means and Standard Deviations of the Kappas for Each Level of the Independent Variable

Group	Size	Mean	Standard Deviation
1, Parallel Forms	20	.3815	.1174
2, KR-20	20	.4750	.1036
3, Kappa	20	.5010	.1297

In order to detect which of the group means was producing the significant results, the Tukey method for comparing the differences between means was used. The results are shown in Table 6.

Table 6

Results of the Tukey Post Hoc Method of Comparing the Differences Between the Kappa Means for the Three Groups

Population Contrast	Sample Constrast	Confidence Interval*	Significant
$\mu_{1,\kappa} - \mu_{2,\kappa}$	-.0935	-.0039, - .1831	yes
$\mu_{1,\kappa} - \mu_{3,\kappa}$	-.1195	-.0300, - .2091	yes
$\mu_{2,\kappa} - \mu_{3,\kappa}$	-.0260	-.1156, + .0636	no

*$1 - \alpha\, 9_{j,j(n-1)}(MSw/n) = (3,405)(.0263) = .0896$

Conclusions

Analysis of variance of the mean KR-20's for each group. As shown in Table 2 the difference in the mean KR-20's for the three groups was not found to be significant at the .05 level of significance. One conclusion that is possible is that tests which are reliable according to parallel forms, KR-20 or kappa criteria will produce the same KR-20 estimates of test-retest reliability. The position which states that there may be a need to reinterpret the reliability coefficients for use with criterion-referenced tests would be upheld in the sense that obtained values of KR-20 for criterion-referenced tests may be lower, but the criterion-referenced tests are probably just as reliable as a norm-referenced test reporting a higher value of KR-20. The mean KR-20 of the three groups of reliable tests ranged from .74 to .79 which is somewhat less than one would require for norm-referenced tests.

Analysis of variance of the kappas for each group. As shown in Table 4 the difference of the mean kappas for the three groups was found to be significant at the .05 level. In order to find out where the differences existed, the Tukey Post Hoc method for comparing the differences among means was used. As indicated in Table 5 both the KR-20 sample and the kappa sample differed from the parallel forms sample, but they did not differ from each other. Since these results were more conclusive than those obtained from the ANOVA performed on the mean KR-20's, they will be discussed in more detail.

The careful development of parallel forms according to the criteria set forth by Paul Horst would seem to be effort wasted in the case of short, criterion-referenced tests. For example, if one had used the criteria for parallel forms to be the determination of the reliability of these test

pairs, only 77 of the 136 test (59%) pairs could have been utilized. Further, a user of those parallel tests would have been disappointed in the results, since the consistency of decisions about examinees for those tests was less than that for either the parallel forms sample or the KR-20 sample. These criteria for parallel forms need further investigation with longer, norm-referenced tests to determine their usefulness and also to determine the range of values for each of the criteria which could be interpreted as being "equal."

Two further conclusions, based upon the practical, can be offered. If one has developed randomly parallel criterion-referenced test, i.e., a set of items measuring a specific objective which have been randomly assigned to two or more tests, then the decision to be made is whether to use KR-20 or kappa. KR-20 is the more commonly reported and its qualities are more familiar to researchers. For those reasons, it may be preferred to kappa. However, if, as these data show, and theoreticians have maintained, the lower obtained values of KR-20 need to be reinterpreted in the context of criterion-referenced testing, then KR-20 may hold little advantage over kappa. Although kappa is readily interpretable as the consistency of decisions made across two tests, after the effects of chance have been removed, the research audience needs to develop an intuitive feeling for the meaning of the various values which kappa can acquire. Further research needs to be done in order to establish what values of kappa are desirable.

Recommendations

Further research needs to be done using the KR-20 coefficient computed across the items of two or more randomly parallel tests and the kappa computed for those tests, in order to determine the relationship between KR-20 and kappa. The question to be addressed is, "Does one have fairly good knowledge of the consistency of decisions that will be made across

two or more tests, if one has knowledge of the level of homogeneity (KR-20) of those two tests?"

The other large topic for future research is that of the behavior of KR-20 and kappa under criterion-referenced test conditions. What level of kappa or KR-20 is desirable, in order to say that two tests are reliable in the consistency of decisions that they produce? Since so much more is known about KR-20, it may be that reinterpretation of KR-20 is more practical than the introduction of an entirely new statistic, particularly since KR-20 can be used with one version of a test while kappa requires parallel forms.

References

Bohrnstedt, G. W. Reliability and Validity Assessment in Attitude Measurement. In G. F. Summers (Ed.), Attitude Measurement. Chicago: Rand McNally, 1971.

Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960, 20, 27-46.

Ebel, R. L. Educational Tests: Valid? Biased? Useful? Phi Delta Kappan, 1975, 57, 83-88.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Glaser, R., & Nitko, A. J. Measurement in Learning and Instruction. In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Glass, G. V., & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Hambleton, R. K., & Novick, M. R. Toward and Integration of Theory and Method for Criterion-Referenced Tests. Journal of Educational Measurement, 1973, 12, 159-170.

Harris, C. W. An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests. Journal of Educational Measurement, 1972, 9, 27-29.

Horst, P. Psychological Measurement and Prediction. Belmont, California: Wadsworth, 1966.

Millman, J. Criterion-Referenced Measurement. In W. J. Popham (Ed.), Evaluation in Education. Berkeley: McCutchan, 1974.

Stanley, J. C. Reliability. In R. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Swaminathan, H., Hambleton, R., & Algina, J. Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. Journal of Educational Measurement, 1974, 11, 263-266.